Title	Perceptual evaluation of voice quality change using paired comparison method
Author(s)	Cheung, Yik-san; 張奕新
Citation	Cheung, Y. [張奕新]. (2014). Perceptual evaluation of voice quality change using paired comparison method. (Thesis). University of Hong Kong, Pokfulam, Hong Kong SAR.
Issued Date	2014
URL	http://hdl.handle.net/10722/238943
Rights	This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.; The author retains all proprietary rights, (such as patent rights) and the right to use in future works.

Perceptual Evaluation Of Voice Quality Change Using Paired Comparison Method

Cheung Yik San

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences, The University of Hong Kong, June 30, 2014)

Abstract

This study investigated the inter-rater and intra-rater reliability of ratings in evaluating voice quality change of a patient using a paired comparison method with a seven-point equal appearing interval scale. Thirty-one naïve listeners, who had no prior perceptual training, and three expert listeners, who had at least five years experience in perceptual voice evaluation, completed a perceptual rating task using the paired comparison method. Results showed that expert listeners achieved a moderate inter-rater reliability (ICC = 0.529) and naïve listeners achieved a fair inter-rater reliability (ICC = 0.347). Intra-rater reliability for expert listeners was significantly higher than naïve listeners (U = 10.0, z = - 2.22, $p < .05$, r = − 0.380). The findings indicated that paired comparison method could be a reliable method for expert listeners in detecting perceptual voice quality change. However, naïve listeners who had no previous perceptual training, may not give as reliable ratings using paired comparison method.

Keywords: perceptual voice evaluation, paired comparison, naïve listeners, expert listeners

Perceptual Evaluation Of Voice Quality Change Using Paired Comparison Method

Perceptual voice rating is used widely in evaluating voice quality (Carding, Carlson, Epstein, Mathieson, & Shewell, 2000). This skill has been considered as an important skill for speech and language pathologists (Carding et al., 2000). Information on voice quality evaluated using perceptual voice evaluation is often considered as a gold standard for comparison with other clinical voice assessment methods, such as acoustic analysis of voice and aerodynamics assessment (de Krom, 1995; Eskenazi, Childers, & Hicks, 1990; Mehta & Hillman, 2008). Despite the frequent use of perceptual voice evaluation in clinical context, the reliability and validity of various kinds of perceptual voice evaluation have been questioned and investigated in recent decades by a number of researchers (Eddins & Shrisvastav, 2013; Kreiman & Gerratt, 1998; Munoz, Mendoza, Fresneda, Carballo, & Ramirez, 2002; Webb et al., 2004).

Limitations of Perceptual Voice Evaluation

The primary limitation of perceptual voice evaluation lied on its varying reliability. Factors affecting reliability of a perceptual voice evaluation included different types of perceptual rating method used (Eddins & Shrivastav, 2013; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Patel, Shrisvastav, & Eddins, 2009; Yiu & Ng, 2004), the use of anchors in rating methods (Chan & Yiu, 2002; Patel, Shrivastav, & Eddins, 2012b), the perceptual parameter measured (Karnell et al., 2007; Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009) or the background and experience of listeners (Kreiman, Gerratt & Ito, 2007; Millet & Dejonckere, 1998). Thus, for perceptual voice evaluation to be used, the reliability of any specific procedure should be investigated.

Perceptual Rating Methods

There are a number of perceptual rating methods reported in the literature. These included visual analogue (VA) scale (Chan & Yiu, 2002), equal appearing interval (EAI) rating scale (Yiu & Ng, 2004), direct magnitude estimation (Patel, Shrivastav, & Eddins, 2009) and matching task (Patel et al., 2009; Patel, Shrivastav & Eddins, 2012a).

Rating scale tasks. Rating scale task is a method that requires listeners to rate the voice quality of a voice on a scale. The scale could be either equal appearing interval (EAI) scale or visual analogue (VA) scale. In perceptual voice evaluation, EAI scale is often used in rating voice qualities such as roughness, breathiness and overall grade of voice quality because of higher reliability than VA scale in the measurement of these metathetic voice qualities (Shrivastav, 2006; Wolf, Martin, & Palmer, 2000; Yiu, Chan, & Mok, 2007) while VA scale was more sensitive in measuring voice signals that are addictive in nature such as loudness and nasality (Stevens, 1975). Thus, when measuring metathetic perceptual parameters such as overall voice quality, the EAI scale would be hypothetically preferred over VA scale. This EAI scale could also be incorporated into other rating methods. For instance, it has been used in a paired comparison paradigm for detection of extent in perceptual difference (Yiu et al., 2007).

Direct magnitude estimation. Direct magnitude estimation, as compared with rating scale task, requires listeners to judge their sensation of voice quality using a ratio with virtually no upper limit in describing the magnitude. However, the assignment of a rating number is of random nature (Eddins & Shrivastav, 2013) and there is no reference point for the judgment of magnitude (Patel et al., 2009; Shrivastav, 2006). Therefore, similar to the VA scale, direct magnitude estimation is a type of magnitude scaling method that is less preferred for measuring metathetic voice qualities (Yiu &

Ng, 2004).

Matching task – use of matching anchor. Matching task is different from the rating-scale and DME tasks in its nature that listeners would manipulate the parameters in a synthesizer until the manipulated signal matches the voice quality of the target stimulus (Kreiman & Gerratt, 2000). There are a number of drawbacks of the matching technique. First, no single standard reference could be used for every matching task (Patel, Eddins & Shrivastav, 2012b). The use of different reference signals in different matching tasks makes direct comparison difficult. Also, it requires a great amount of time for completion on average as compared with rating scale task and DME (Patel et al., 2009). The use of anchor was considered to be capable of eliminating unstable internal representation, hence improving the reliability of ratings (Gerratt & Kreiman, 2001). Such use of anchor in perceptual voice evaluation has been supported by the literatures (Chan & Yiu, 2002; Kreiman & Gerratt, 2000; Patel et al., 2012a; Yiu et al., 2007) and has been further developed and adopted in other rating method.

Paired comparison task – use of comparison anchor. One of the methods adopting the use of anchor was the paired comparison task. Unlike the matching task, the comparison in a paired comparison task is to "compare" the difference between two stimuli (Yiu et al., 2007). In practice, a listener would listen to and compare a pair of voice samples. And the listener has to rate the second sample with reference to the first sample.

The four auditory perceptual voice evaluation tasks use very different methodologies. While rating scale task and direct magnitude estimation were more suitable in rating addictive voice qualities such as nasality (Stevens, 1975), the matching tasks and paired comparison tasks were more sensitive in measuring metathetic qualities such as hoarseness and overall grade of voice quality (Stevens,

1975). As this study focused on overall voice quality, which is a metathetic voice quality, the VA scale and the direct magnitude estimation tasks were therefore not chosen. As compared to the matching task, paired comparison task was a relatively simple evaluation procedure that could be applied to clinical setting without using additional anchors, whether synthesized or natural anchors (Kreiman & Gerratt, 2000; Patel et al., 2012b). It was therefore adopted in this study to investigate voice quality change.

Different Anchors In Paired Comparison Task

As mentioned earlier, there is a need to investigate the reliability of any specific procedure in perceptual voice evaluation. Literatures have investigated the reliability of different paired comparison tasks. For instance, in the study by Chan and Yiu (2006), listeners were asked to judge whether a pair of synthesized voices were identical or different in breathiness severity. Their task only required the listeners to respond on a binary choice. Such task was found to be effective in improving naïve listeners' reliability in perceptual rating. However, as no rating scale was adopted in the study, the extent of difference between the anchor and stimuli could not be established. Yiu et al. (2007) attempted to explore this gap in literature and investigated the validity of a paired comparison paradigm with the use of an EAI rating scale. The paradigm focused on direct comparison of difference between anchor and stimulus by asking the listeners to rate the extent of difference perceived in a comparison pair. Result showed that this paired comparison paradigm was effective in improving the inter-rater reliability significantly ($F(1,99) = 4.90$, $p < .05$)). In the study by Chan and Yiu (2002), the effect of synthesized and natural anchors on reliability of perceptual voice evaluation was also investigated. Inter-rater agreement for synthesized anchors was found to be better than natural anchors. However, the study

has a number of limitations. The natural anchors used were voices from different dysphonic patients so the judgment was made on between-subject difference. In fact, natural anchors could also be selected from different time of intervention of the same patient. The reliability of perceptual ratings judged on such within-subject comparison was not investigated. Such within-subject comparison is of high clinical relevance because it could be used to detect the change of voice quality after voice therapy. In light of this clinical relevance, this study sought to investigate the reliability of perceptual voice evaluation in judging voice samples collected from the same person. Paired comparison paradigm was employed for the present study because of its intuitive simplicity when compared with the other ratings methods (Patel et al., 2009; Yiu et al., 2007); and the concept of the use of anchor, i.e. one stimulus acts as the anchor for the second stimulus to be rated (Kreiman et al., 2007; Yiu et al., 2007). Based on the paradigm proposed by Yiu et al. (2007), a seven-point EAI scale from -3 to +3 was used in this study with the negative ratings representing regression while positive ratings indicating improvement. The anchor and stimuli in this paradigm were voices of dysphonic patients recorded at different time of intervention. Comparison made within each pair reflected the listeners' perception of subtle change in the overall voice quality of the same dysphonic patient over time.

Apart from the rating method, the type of voice quality to be measured would be another factor to be considered.

Perceptual Parameter To Be Measured

It is important to measure a clinically meaningful perceptual parameter in perceptual voice evaluation (Kempster et al., 2009). In perceptual voice evaluation, listeners would subjectively rate perceptual parameter(s). Rating is given either by a) rating several perceptual parameters or b) rating only one specific perceptual parameter.

For the first rating method, several protocols are available. For instance, GRBAS scale proposed by the Japanese Society of Logopedics and Phoniatrics and the Vocal Profile Analysis Scheme (Laver, Wirz, MacKenzie, & Hiller, 1981). However, using these protocols to rate a large number of voice samples would be too time-consuming (Webb et al., 2004). A more practical and faster way would be to rate one specific representative perceptual parameter only. The literatures have reports that examined the validity and reliability of perceptual voice evaluation in rating specific voice qualities (Millet & Dejonckere, 1998; Shrivastav, 2006; Webb et al., 2004; Wolfe et al., 2000). Most of the studies support the use of overall grade of voice quality as the voice quality measurement to be used (Webb et al., 2004; Yamaguchi, Shrivastav, Andrews, & Nimii, 2003), which is defined by "overall impression of the deviance in voice quality" (Millet & Dejonckere, 1998). The validity and reliability was found to be high for this perceptual parameter as supported by Millet and Dejonckere (1998) that overall grade could be rated more reliably (Spearman's Rho = 0.86) than the degree of breathiness and asthenia (Spearman's Rho = 0.64) among the five voice qualities in GRBAS scale. Yamaguchi et al. (2003) found similar results and reported that inter-rater reliability to be high for the "overall grade" (mean inter-rater correlation = 0.893) in their study investigating perceptual ratings given by listeners of different nationality. According to the literature reviewed above, overall grade in voice quality is a valid and reliable perceptual parameter to be rated and it was adopted in this study.

Background and Experience of Listeners

Another factor affecting the reliability of a perceptual voice evaluation is the experience of the listeners. It has been shown that different listener groups would demonstrate different reliability in perceptual ratings (Lee, Drinnan & Cardings, 2005; Munoz et al., 2002). Lee et al. (2005) investigated the perceptual voice evaluation of

dysphonic individuals with their voice using the GRBAS protocol. The reliability of rating was measured by test-retest agreement using weighed Kappa statistics. It was found that these naïve listeners demonstrated relatively low intra-rater reliability (Kappa = 0.51) for G (overall grade in voice) than the ratings given by expert clinicians (Kappa = 0.74). Helou et al. (2010) found similar findings on the study of perceptual ratings given by different listeners using CAPE-V protocol. They also found the naïve listeners had lower intra-rater agreement and inter-rater agreement (r = 0.528) than expert listeners including Ear, Nose and Throat Specialists and speech therapists (r = 0.722). It has been suggested that experience in perceptual voice evaluation allows expert listeners to have a more stable internal representation of voice qualities (Kreiman et al., 2007) than naïve listeners. Apart from experience, more stable internal representation could also be established by providing perceptual training (Kreiman et al., 2007). The effect of training on perceptual voice evaluation has been shown to be beneficial (Chan & Yiu, 2002; Chan & Yiu, 2006; Millet & Dejonckere, 1998). Inter-rater reliability and intra-rater reliability of ratings were both increased due to training effects ($F(1,27) = 5.70$, $p < .05$) after training on perceptual voice evaluation (Chan & Yiu, 2002).

This study aimed to investigate whether there was a difference in detecting voice quality between two samples by naïve and expert listeners. It was hypothesized that expert listeners would give higher inter-rater and intra-rater reliability under this paired comparison method as their experience and training established more stable internal representation than naïve listeners.

Method

Participants

Two different groups of listeners recruited in this study were: a) naïve listeners,

who had not received any perceptual training, and b) expert listeners, who had experience in at least five years as clinicians. 31 post-secondary students (who had no auditory training before) and three expert listeners (One professor, one associate professor and one assistant professor from the Division of Speech and Hearing Sciences, The University of Hong Kong) completed written consent forms and agreed to participate in this study. A perceptual rating task was conducted in which every listener rated all the voice samples.

Stimuli

Stimuli were voice samples selected from a double-blinded randomized control study by Yiu et al. (to be submitted) relating to acupuncture treatment for voice problems. Voices samples from 40 subjects were collected before (Pre) and after (Post1 and Post2) the acupuncture treatment session. The target sentence recorded in each stimulus was a Cantonese sentence /pa1 pa1 ta2 ko1 ko1/ (爸爸打哥哥) and each subject produced the same sentence repeatedly three times during each recording. For the paired comparison approach adopted in this present study, every “pair” was comprised of two voices from each patient. Three pairs, respectively Pre-Pre pair, Pre-Post1 pair and Pre-Post2 pair, were created from each patient. Half of the stimuli were arranged in reverted order (Post2-Pre pair and Post1-Pre pair) to control the presence of practice effects when listeners were rating the voice samples.

Procedure

Every participant in this study was asked to complete a perceptual rating task. The perceptual rating task for the naïve listener was conducted in sound-treated rooms at the Division of Speech and Hearing Sciences, University of Hong Kong. Voice stimuli were presented through an Apple MacBook Pro (mid 2012) using a headphone output (Hyundai HY 7557). The other three expert listeners completed the evaluation

task at their office with their preferred headphones.

There were three practice trials and 135 experiment trials. The listeners evaluated a pair of voice stimuli in every trial. By comparing to the first stimulus (anchor), they rated the overall grade of voice quality of the second stimulus on a seven-point EAI scale from -3 (most negative change) to +3 (most positive improvement), with 0 representing no change. They were instructed to focus on the overall voice quality they perceive. Three practicing trials were conducted first for them to get familiarized with the task.

Each voice pair was presented using a Microsoft PowerPoint slideshow. All slideshows were randomized in order. Over ten percent of the pairs, i.e. 15 pairs, were repeatedly presented to investigate the intra-rater reliability. For every 45 slides listened, listeners were given a one-minute break.

The ratings given by the listeners for Post1-Pre pairs and Post2-Pre pairs were converted by multiplying with -1 so that the ratings were considered to be referencing to the “Pre” stimuli (anchors). For example, a score of +3 rated for Post2-Pre pair was converted to -3 as if the set was seen as a Pre-Post2 pair. The ratings given by the listeners for Pre-Pre pairs, Pre-Post1 pairs and Pre-Post2 pairs were not adjusted at all.

Data Analysis

To address the research question, both inter-rater reliability and intra-rater reliability were determined.

Inter-rater reliability for naïve listener group and expert group were assessed using intraclass correlation coefficient (ICC). ICC is an index commonly used to evaluate the consistency in measurement made by more than two listeners and it represents the ratio of between-subject variance to total variance (Shrout & Fleiss, 1979; Tinsley & Brown, 2000). It was chosen because it could quantify the extent to which

individuals' ratings resemble each other and could effectively determine the inter-rater reliability, as it takes chance level into account (Field, 2005). A two-way random effects model with the specification of ICC for absolute agreement was used for this study as suggested by Field (2005).

Intra-rater reliability for every listener in naïve group and expert group was assessed using percent agreement and ICC. Percent agreement could evaluate the degree to which the ratings are identical in test and retest trials (Gisev, Bell & Chen, 2013). One advantage of percentage agreement is its simple calculation, and it could be used for assessing reliability of any number of listeners. However, it does not account for the agreement achieved by chance (Lombard, Snyder-Duch, & Bracken, 2002) whereas ICC does. Both ICC and percent agreement were therefore used to investigate intra-rater reliability comprehensively.

Results

In order to investigate the general characteristics of ratings, the summary of all ratings given by both naïve listeners and expert listeners were calculated using descriptive statistics. Table 1 shows the descriptive statistic analysis for the mean value, maximum and minimum values, and range of the ratings on overall voice quality change rated by each naïve listeners. Table 2 lists the descriptive statistics for ratings given by each expert listener.

Table 1

Descriptive Statistics of Ratings By Each Naïve Listener

	Ratings given				
	Mean	SD	Range	Minimum	Maximum
Naive1	0.27	1.13	5.00	-2.00	3.00
Naive2	0.09	1.09	4.00	-2.00	2.00
Naive3	0.22	0.97	5.00	-2.00	3.00
Naive4	0.31	1.12	6.00	-3.00	3.00
Naive5	0.09	1.24	6.00	-3.00	3.00
Naive6	0.15	0.71	4.00	-2.00	2.00
Naive7	0.39	1.42	6.00	-3.00	3.00
Naive8	0.36	1.16	5.00	-2.00	3.00
Naive9	0.13	0.82	4.00	-2.00	2.00
Naive10	0.24	1.00	5.00	-2.00	3.00
Naive11	0.29	1.06	5.00	-2.00	3.00
Naive12	0.21	0.95	4.00	-2.00	2.00
Naive13	0.16	1.34	6.00	-3.00	3.00
Naive14	0.07	1.22	6.00	-3.00	3.00
Naive15	0.13	1.05	4.00	-2.00	2.00
Naive16	0.26	1.26	5.00	-2.00	3.00
Naive17	0.23	1.27	6.00	-3.00	3.00
Naive18	0.22	1.17	5.00	-2.00	3.00
Naive19	0.24	1.03	6.00	-3.00	3.00
Naive20	0.08	1.23	6.00	-3.00	3.00
Naive21	0.22	1.17	6.00	-3.00	3.00
Naive22	0.26	0.73	4.00	-2.00	2.00
Naive23	0.12	1.27	6.00	-3.00	3.00
Naive24	0.13	1.16	4.00	-2.00	2.00
Naive25	0.30	1.23	6.00	-3.00	3.00
Naive26	0.24	0.89	5.00	-2.00	3.00
Naive27	0.16	0.87	4.00	-2.00	2.00
Naive28	0.27	1.33	5.00	-2.00	3.00
Naive29	0.30	1.09	5.00	-2.00	3.00
Naive30	0.15	1.00	4.00	-2.00	2.00
Naive31	0.10	1.09	5.00	-3.00	2.00
Combined	0.21	1.10	5.06	-2.39	2.68

Table 2

Descriptive Statistics of Ratings By Each Expert Listener

	Ratings given on overall voice quality change				
	Mean	SD	Range	Minimum	Maximum
Expert1	0.31	0.96	5.00	-2.00	3.00
Expert2	0.31	1.03	5.00	-2.00	3.00
Expert3	0.39	1.09	6.00	-3.00	3.00
Combined	0.34	1.03	5.33	-2.33	3.00

From descriptive statistics, it could be seen that both groups were able to identify a difference in voice samples to a certain extent from the means of their ratings. Expert listeners gave generally higher ratings (Mean = 0.34) than naïve listeners (Mean = 0.21).

Also, expert listeners gave a wider range in ratings (Mean range = 5.33) than naïve listeners (Mean range = 5.06). Expert listeners were able to rate without much deviation from mean (SD = 1.03) while ratings by naïve listeners were more deviated from mean (SD = 1.10).

Since expert listeners gave higher ratings than naïve listeners, such difference in ratings was further investigated. Due to the small sample size for expert group (three expert listeners), the assumption of normal distribution was violated. Therefore, a non-parametric Mann-Whitney *U* test was selected for the analysis of the difference in ratings across the two groups.

Result from Mann-Whitney *U* test shows that ratings by expert listeners (Median = 0.311) were significantly higher than ratings given by naïve listeners (Median = 0.222), $U = 5.5$, $z = -2.492$, $p < .05$, $r = -.427$. The effect size - 0.427 corresponds to a medium to large effect size (McKnight & Najab, 2010).

In the determination of inter-rater reliability, the value of intraclass correlation coefficient was calculated using a two-way random effects model. Table 3 shows the inter-rater reliability of ratings by the two groups.

Table 3

Inter-Rater Reliability of Ratings By Naïve And Expert Listeners

	Inter-rater reliability	*p* value	95% confidence interval
Naïve listeners	ICC (3,1) = 0.347	$p < .01$	0.294 – 0.411
Expert listeners	ICC (3,1) = 0.529	$p < .01$	0.431 – 0.620

ICC = Intraclass correlation coefficient

Naïve listeners had poorer inter-rater reliability than expert listeners. The value of intraclass correlation coefficient for naïve listeners was 0.347. The value corresponds to “fair” reliability according to Landis and Koch (1977). ICC value for expert group was 0.529, which corresponds to “moderate reliability”.

Apart from inter-rater reliability, intra-rater reliability was calculated using test-retest percent agreement on 10% of all the items and also using ICC. Table 4 and table 5 show the percentage of agreement within 1-point, percentage of exact agreement and the ICC value for naive listeners and expert listeners respectively.

Table 4

Intra-Rater Reliability of Naïve Listeners

Naïve Listener	Percent agreement		ICC (3,1)
	One-point agreement	Exact agreement	
listener1	80%	60%	0.542*
listener2	87%	33%	0.559*
listener3	73%	40%	0.010*
listener4	93%	53%	0.715**
listener5	80%	20%	0.207
listener6	94%	80%	0.697**
listener7	67%	20%	0.041
listener8	100%	27%	0.737**
listener9	87%	67%	0.497*
listener10	87%	53%	0.351
listener11	87%	40%	0.547*
listener12	100%	53%	0.779**
listener13	87%	27%	0.395
listener14	73%	60%	0.092
listener15	73%	33%	0.247
listener16	87%	47%	0.751**
listener17	67%	53%	0.125
listener18	73%	40%	0.391
listener19	93%	27%	0.295
listener20	60%	20%	0.287
listener21	73%	20%	0.263
listener22	100%	80%	0.732**
listener23	80%	27%	0.413*
listener24	73%	53%	0.512*
listener25	73%	20%	0.240
listener26	87%	53%	0.497*
listener27	73%	47%	0.208
listener28	80%	40%	0.510*
listener29	80%	53%	0.508*
listener30	73%	40%	0.138
listener31	73%	27%	0.332
Mean	81%	42%	0.407
Range	67% - 100%	20% - 80%	0.01 - 0.779

Note. *$p < .05$, **$p < .01$

Table 5

Intra-Rater Reliability of Expert Listeners

	Percent agreement		
Expert Listener	One-point agreement	Exact agreement	ICC (3,1)
listener1	100%	73%	0.901**
listener2	93%	53%	0.704**
listener3	93%	73%	0.698**
Mean	95.3%	67%	0.768
Range	93% - 100%	53% - 73%	0.698 – 0.901

Note. **$p <$ .01

From table 4, the percent agreement for naïve listeners was moderately high. The exact agreement was 42% and the percent agreement within one-point was 81%. Naïve listeners had a wide range in intra-rater reliability in terms of exact agreement, ranging from 20% to 80%. With chance level corrected, ICC value for naïve listeners is 0.407, representing a fair reliability. The range of ICC value was large, ranging from 0.01 to 0.779.

From table 5, the percent agreement for expert listeners was high. The exact agreement was 67% and the percent agreement within one-point was 95.3%. Expert listeners had a narrow range in intra-rater reliability in terms of exact agreement, ranging from 53% to 73%. With chance level corrected, ICC value for expert listeners is 0.768, representing a substantial reliability. The range of ICC value was narrow, ranging from 0.698 to 0.901.

The average ICC value for intra-rater reliability of expert listeners was found to be higher than that of naïve listeners. Therefore, a non-parametric Mann-Whitney *U* test was used to investigate the difference. Result from Mann-Whitney *U* test shows that the intra-rater ICC of expert listeners (Median = 0.704) was significantly

higher than that of the intra-rater ICC of naïve listeners (Median = 0.395), $U = 10.0$, z = - 0.22, $p < .05$, $r = -.380$. The effect size - 0.380 corresponds to a medium to large effect size (McKnight & Najab, 2010).

Discussion

The objective of the present study was to investigate the reliability of ratings given by naïve and expert listeners on the perceptual evaluation of voice quality change using a paired comparison paradigm. Our result showed that both naïve and expert listeners were able to detect a subtle change in voice quality and expert listeners had better reliability than naïve listeners.

Perceptual Evaluation of Voice Quality Change

Both naïve listeners and expert listeners were able to evaluate a change in voice quality. The mean range of rating was 5.33 for expert listeners and 5.06 for naïve listeners on the seven-point EAI scale adopted in this study. Generally speaking, both groups rated a slight improvement on the overall voice quality change according to their mean ratings. This indicates that under the use of paired comparison method, both naïve listeners and expert listeners were able to detect a subtle change in voice quality of the same dysphonic patient. This result shows that when natural voices of the same dysphonic patient are chosen as anchor and stimulus, detection of subtle change in voice quality is possible. While the two listeners groups were able to detect a difference in voice quality, expert listeners rated voice samples with higher ratings than naïve listeners according to result from Mann-Whitney *U* test. This is very likely to be attributed to expert listeners' experience as they have better internal representation of voice quality standards and could rate any subtle changes more sensitively than naïve listeners. The findings from this study are important as they contrast to the results by Chan & Yiu (2002) in which reliability in

rating for natural anchor and stimulus selected from different dysphonic patients was lower than that of synthesized anchors. This adds insights to future directions of clinical application of paired comparison method in rating voice quality change of the same patient, as this method is more convenient than synthesizing artificial anchors for every patient.

Paired Comparison Method: Rater Reliability

The intra-rater reliability in this study was relatively high for both naïve listeners and expert listeners when rating voice quality change. Intra-rater reliability using percent agreement reached a level as high as 81% and 95.3% for naïve listeners and expert listeners respectively. It is true that percent agreement may not account for the agreement by chance level. However, the intra-rater reliability was still concluded to be high according to the significantly higher ICC values of expert listeners. The finding on the high intra-rater reliability is a piece of supportive evidence to the application of paired comparison method in evaluating subtle change in voice quality for expert listeners. In this study, the intra-rater reliability for expert listeners (ICC = 0.768) is comparable to findings from recent studies on perceptual evaluation, such as a recent study by Law et al. (2010) for which intra-rater ICC value was 0.73 for rating sustained vowel and 0.828 for rating passage using non-anchored perceptual method; and a study by Helou et al. (2010) for which intra-rater ICC value was 0.911 using CAPE-V to rate postthyroidectomy voices. Apart from these two studies, Yiu et al. (2007) reported a 51.5% in intra-rater exact agreement using synthesized anchors in a paired comparison method. The intra-rater exact agreement for expert listeners in this study reached 67%, which is higher than that reported by Yiu et al. (2007). Comparison of findings between this study and other perceptual studies shows that the paired comparison method adopted in this study is likely to be a reliable

perceptual method in rating subtle difference between voice samples of the same patient. This finding is very encouraging because our result is in line with the widely supported use of anchor (Kreiman & Gerratt, 2000; Patel et al., 2012b; Yiu et al., 2007) in the paired comparison, which is believed to be able to replace unstable internal standard with stable external representation (Kreiman et al., 2007). The high intra-rater reliability, especially for expert listeners, is likely to be resulted from the use of anchor.

Rater Experience

Another finding from this study is the difference in reliability of ratings by the two different listener groups. Inter-rater reliability for expert listeners was moderate (ICC = 0.529) whereas for naïve listeners, inter-rater reliability only reached fair level (ICC = 0.347). Intra-rater rater reliability of expert listeners (ICC = 0.768) was found to be significantly higher than naïve listeners (ICC = 0.407) as well according to Mann-Whitney *U* test ($U = 10.0$, $z = -2.22$, $p < .05$). These findings are suggestive of better reliability by expert listeners than naïve listeners when using a paired comparison approach. It is concordant to the results from other perceptual voice evaluation studies (Helou et al., 2010; Lee et al., 2005) for which expert listeners gave better reliability than non-experts. Such concordance could be attributed to the rater experience in perceptual rating. The three expert listeners recruited in the study had at least five years of clinical and research experience in perceptual rating studies. On the other hand, the naïve listeners had no prior training in perceptual analysis. As a result, a difference between their perceptual rating experiences exists. For the expert listeners, it is likely that they would have more stable internal standards of the voice qualities to be rated as benefited from their experiences (Kreiman, 2007). On the other hand, naïve listeners have a less stable representation of internal standards. So,

when rating a voice quality change using perceptual parameter as simple as “overall grade of voice quality”, naïve listeners may not be able to give adequately sensitive ratings towards a subtle voice quality change. This finding shows that even for some very simple perceptual evaluation methods, such as the paired comparison method used in this study, other complementary trainings may still be necessary for naïve listeners to give reliable ratings.

Perceptual Training

Despite the simplistic procedure used in paired comparison in this study, the lack of auditory training could explain the relatively lower inter-rater reliability of naïve listeners. Literature reported that training is beneficial to perceptual voice evaluation (Chan & Yiu, 2002; Chan & Yiu, 2006; Millet & Dejonckere, 1998). These trainings could help build a stable internal representation of voice qualities hence comparison between voices of the same patient would be made more reliable. Naïve listeners in the study lacked this training beforehand. In this study, they were only allowed three practice trials to familiarize with perceptual rating task without structured perceptual training. Therefore, the lack of training may be one contributing factor to explain their lower intra-rater and lower inter-rater reliability. This finding encourages further investigation on whether training using this paired comparison method would increase reliability of ratings by naïve listeners.

Stimuli

This study employed stimuli from previous study by Yiu et al. (to be submitted). These stimuli were voices of dysphonic patients receiving acupuncture treatment. In general, post1 and post2 voices were believed to have improvement in overall voice quality after acupuncture treatment according to the results of that study. However, the objective of this study is to investigate whether naïve and expert

listeners could perceive the subtle difference between two stimuli from the same patient only. The nature of the voice samples, whenever the time of intervention or whatever the type of treatment, did not confound to the objective of this study. Furthermore, the order of presentation of stimuli sets was randomized. Therefore the confounding practice effect was controlled.

However, the stimuli used in this study could be somewhat distracting to naïve listeners. The recorded voice from Yiu's study was "/pa1 pa1 ta2 ko1 ko1/ - /pa1 pa1 ta2 ko1 ko1/ - /pa1 pa1 ta2 ko1 ko1/" (爸爸打哥哥 -爸爸打哥哥 -爸爸打哥哥). For every set, listeners would listen to six /pa1 pa1 ta2 ko1 ko1/ sentences for comparison of voice quality change. The repetitive nature of the testing stimuli could disturb the internal representation of voice qualities for naïve listeners, or even expert listeners. Despite the instruction to focus on overall grade of voice quality only, such repetition could still have affected their validity in comparison, which may explain the lower inter-rater reliability for naïve listeners. A possible conclusion drawn would be that external representation from anchor, when too lengthy, may impose disturbance towards internal representation, resulting in less reliable ratings as reflected from this study. A focused single-sentence comparison could be adopted in future when using paired comparison for comparing subtle voice quality difference. Multiple repetitions of sentence stimuli shall be avoided.

Rating Scale

The EAI scale used in this study was comprised of seven equidistant points, with three points diverging from center zero. The narrowness of the seven-point scale may have caused reduction in sensitivity in a way that listeners tend to restrict ratings towards central point. In this study, the ratings given by both listener groups showed a tendency towards the center of scale. This could be explained by the relatively low

sensitivity of a narrow scale. In order to avoid such central tendency effect that hinders the sensitivity of the scale, a longer-ranged scale, such as 11-point scale, could be used instead. In this way, ratings would be more likely to spread apart from center point and sensitivity towards subtle voice quality change could also be increased.

Limitations And Future Directions

Some practical limitations have been noted in this study. First, expert listeners in the study used their private preferred headphones while naïve listeners used a high-quality headphone. The degree of the effect of different filters on different headphones contributing towards statistical results was not known. Also, the number of stimuli presented to listeners was large. A total number of 828 "/pa1 pa1 ta2 ko1 ko1/ (爸爸打哥哥)" voice segments were presented to listeners within one hour. To naïve listeners, such large load in auditory perception could impose fatigue and affect their auditory attention. The sample size of listeners was not large enough and may not represent population of naïve and expert listeners.

This study has led to some possible directions for future clinical and research application. First, it encourages the use of paired comparison approach in perceiving voice quality change of one patient. Expert listeners, who had prior perceptual training and experience in perceptual voice evaluation, might use this method for rating quality change of patients' voices. Second, whether paired comparison may be considered a suitable method for perceptual training or not could be set as further research objective. This study also encourages similar future study on detection of different dysphonic patients' voice quality change using different perceptual parameters as outcome measure. For instance, breathiness could be targeted for patients with vocal fold pathologies using similar paired comparison framework.

Conclusion

To conclude, the results from this study showed that expert listeners could give reliable ratings on the detection of voice quality change of a patient using a paired comparison approach. While naïve listeners had good test-retest intra-rater agreement, their inter-rater reliability of ratings was lower than that by expert listeners. This finding is consistent with results from other perceptual studies supporting the use of perceptual training for naïve listeners. This study also suggested the use of longer range EAI scale to enhance the sensitivity of ratings in perceptual voice evaluation. Finally, this study lays ground for possible future application of paired comparison method in clinical setting when rating voice quality change of dysphonic patient, such as voice recorded at different time of intervention.

Acknowledgement

The author would like to express immense gratitude to Professor Edwin Yiu for his valuable guidance, professional advices and unlimited support in this dissertation. This dissertation would not have been successful without his supervision. The author would also like to thank Dr. Karen Chan and Dr. Estella Ma for their very kind support and participation in this study.

The author would like to express special thanks to Miss Nikki Sung Hei Tung for her unconditional mental support and all-rounded care throughout the course of this dissertation, especially at hard times. This dissertation would not have been possibly accomplished without her full backup.

References

Carding, P., Carlson, E., Epstein, R., Mathieson, L., & Shewell, C. (2000). Formal perceptual evaluation of voice quality in the United Kingdom. *Logopedics, Phoniatrics and Vocology, 25*, 133-138.

Chan, K. M., & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language & Hearing Research*, *45*(1).

Chan, K. M., & Yiu, E. M. (2006). A comparison of two perceptual voice evaluation training programs for naive listeners. *Journal of Voice, 20*(2), 229-241.

de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38,* 794–811.

Eddins, D. A., & Shrivastav, R. (2013). Psychometric properties associated with perceived vocal roughness using a matching task. *The Journal of the Acoustical Society of America*, *134*(4), *EL294-EL300*.

Eskenazi, L., Childers, D. G., Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33, 298-306.*

Field, A. P. (2005). Intraclass correlation. *Encyclopedia of statistics in behavioral science*.

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America*, *110*, 2560.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Inter-rater agreement and inter-rater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Helou, L. B., Solomon, N. P., Henry, L. R., Coppit, G. L., Howard, R. S., &

Stojadinovic, A. (2010). The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. *American Journal of Speech-Language Pathology, 19*(3), 248-258.

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21*(5), 576-590.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol. *American Journal of Speech-Language Pathology, 18*(2), 124-132.

Kreiman, J. & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. Journal of the acoustical society of America, *104*, 1598-1608.

Kreiman, J., & Gerratt, B. R. (2000). Measuring vocal quality. *Voice quality measurement*, 73-102.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech, Language and Hearing Research, 36*(1), 21.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*, *122*, 2354.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, *33*(1), 159-174.

Laver, J., Wirz, S., MacKenzie, J., & Hiller, H. (1981) *A perceptual protocol for the analysis of vocal profiles.* University of Edinburgh.

Lee, M., Drinnan, M., & Carding, P. (2005). The reliability and validity of patient self-rating

of their own voice quality. *Clinical Otolaryngology*, *30*(4), 357-361.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability.*Human communication research*, *28*(4), 587-604.

McKnight, P. E., & Najab, J. (2010). Mann‐Whitney *U* Test. *Corsini Encyclopedia of Psychology*.

Mehta, D. D., & Hillman, R. E. (2008). Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Current opinion in otolaryngology & head and neck surgery*, *16*(3), 211.

Millet, B., & Dejonckere, P. H. (1998). What determines the differences in perceptual rating of dysphonia between experienced rater? *Folia Phoniatrica et Logopedica, 50*, 305-310.

Munoz, J., Mendoza, E., Fresneda, M. D., Carballo, G., & Ramirez, I. (2002). Perceptual analysis in different voice samples: agreement and reliability. *Perceptual and motor skills*, *94*(3c), 1187-1195.

Patel, S., Shrivastav, R., & Eddins, D. A. (2009). Perceptual Distances of Breathy Voice Quality: A Comparison of Psychophysical Methods. *Journal of Voice*, *24*(2), 168-177

Patel, S., Shrivastav, R., & Eddins, D. A. (2012a). Developing a single comparison stimulus for matching breathy voice quality. *Journal of Speech, Language and Hearing Research*, *55*(2), 639.

Patel, S., Shrivastav, R., & Eddins, D. A. (2012b). Identifying a comparison for matching rough voice quality. *Journal of Speech, Language and Hearing Research*, *55*(5), 1407.

Shrivastav, R. (2006). Multidimensional scaling of breathy voice quality: individual

differences in perception. *Journal of Voice, 20*(2), 211-222.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Stevens, S. S. (1975). *Psychophysics*. Transaction Publishers.

Tinsley, H. E., & Brown, S. D. (Eds.). (2000). *Handbook of applied multivariate statistics and mathematical modeling*. Academic Press.

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, *261*(8), 429-434.

Wolfe, V. I., Martin, D. P., & Palmer, C. I. (2000). Perception of dysphonia voice quality by naive listeners. *Journal of Speech, Language, and Hearing Research, 43*(3), 697-705

Yamaguchi, H., Shrivastav, R., Andrews, M. L., & Niimi, S. (2003). A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia phoniatrica et logopaedica*, *55*(3), 147-157.

Yiu, E. M. L., Chan, K. M., & Mok, R. S. M. (2007). Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical linguistics & phonetics*, *21*(2), 129-145.

Yiu, E. M. L., & Ng, C. Y. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical linguistics & phonetics*, *18*(3), 211-229.

Appendix A
Rating forms for perceptual voice evaluation

Perceptual Voice Evaluation – Matched Comparison
EAI Rating form

Name: _____ Age/Sex:_____

Instruction

You are going to listen to 135 sets of voices presented by a Powerpoint slideshow.

Each set contains two voice samples: the **reference** and the **stimuli**.
You are required to **first** carefully listen to the reference once, and **then** listen to the stimuli once.
You may listen to the set of voices again.

After listening to both reference and stimuli, please pay attention to and rate the **overall voice quality of stimuli** as compared to the reference. +3 represents the best improvement as compared to reference, 0 would represent no change between stimuli and reference, and -3 represents the worst regression as compared to reference.

You have to rate by filling the box of the corresponding rating. For instance, if you think there is no change between stimuli as compared to reference, please fill in the box representing 0.

-3	-2	-1	0	+1	+2	+3
☐	☐	☐	■	☐	☐	☐

You would first practice on 3 trials.

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Practice1	☐	☐	☐	☐	☐	☐	☐
Practice2	☐	☐	☐	☐	☐	☐	☐
Practice3	☐	☐	☐	☐	☐	☐	☐

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Set1	☐	☐	☐	☐	☐	☐	☐
Set2	☐	☐	☐	☐	☐	☐	☐
Set3	☐	☐	☐	☐	☐	☐	☐
Set4	☐	☐	☐	☐	☐	☐	☐
Set5	☐	☐	☐	☐	☐	☐	☐
Set6	☐	☐	☐	☐	☐	☐	☐
Set7	☐	☐	☐	☐	☐	☐	☐
Set8	☐	☐	☐	☐	☐	☐	☐
Set9	☐	☐	☐	☐	☐	☐	☐
Set10	☐	☐	☐	☐	☐	☐	☐
Set11	☐	☐	☐	☐	☐	☐	☐
Set12	☐	☐	☐	☐	☐	☐	☐
Set13	☐	☐	☐	☐	☐	☐	☐
Set14	☐	☐	☐	☐	☐	☐	☐
Set15	☐	☐	☐	☐	☐	☐	☐

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Set16	☐	☐	☐	☐	☐	☐	☐
Set17	☐	☐	☐	☐	☐	☐	☐
Set18	☐	☐	☐	☐	☐	☐	☐
Set19	☐	☐	☐	☐	☐	☐	☐
Set20	☐	☐	☐	☐	☐	☐	☐
Set21	☐	☐	☐	☐	☐	☐	☐
Set22	☐	☐	☐	☐	☐	☐	☐
Set23	☐	☐	☐	☐	☐	☐	☐
Set24	☐	☐	☐	☐	☐	☐	☐
Set25	☐	☐	☐	☐	☐	☐	☐
Set26	☐	☐	☐	☐	☐	☐	☐
Set27	☐	☐	☐	☐	☐	☐	☐
Set28	☐	☐	☐	☐	☐	☐	☐
Set29	☐	☐	☐	☐	☐	☐	☐
Set30	☐	☐	☐	☐	☐	☐	☐

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Set31	☐	☐	☐	☐	☐	☐	☐
Set32	☐	☐	☐	☐	☐	☐	☐
Set33	☐	☐	☐	☐	☐	☐	☐
Set34	☐	☐	☐	☐	☐	☐	☐
Set35	☐	☐	☐	☐	☐	☐	☐
Set36	☐	☐	☐	☐	☐	☐	☐
Set37	☐	☐	☐	☐	☐	☐	☐
Set38	☐	☐	☐	☐	☐	☐	☐
Set39	☐	☐	☐	☐	☐	☐	☐
Set40	☐	☐	☐	☐	☐	☐	☐
Set41	☐	☐	☐	☐	☐	☐	☐
Set42	☐	☐	☐	☐	☐	☐	☐
Set43	☐	☐	☐	☐	☐	☐	☐
Set44	☐	☐	☐	☐	☐	☐	☐
Set45	☐	☐	☐	☐	☐	☐	☐

YOU NOW HAVE 1-MINUTE BREAK

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Set46	☐	☐	☐	☐	☐	☐	☐
Set47	☐	☐	☐	☐	☐	☐	☐
Set48	☐	☐	☐	☐	☐	☐	☐
Set49	☐	☐	☐	☐	☐	☐	☐
Set50	☐	☐	☐	☐	☐	☐	☐
Set51	☐	☐	☐	☐	☐	☐	☐
Set52	☐	☐	☐	☐	☐	☐	☐
Set53	☐	☐	☐	☐	☐	☐	☐
Set54	☐	☐	☐	☐	☐	☐	☐
Set55	☐	☐	☐	☐	☐	☐	☐
Set56	☐	☐	☐	☐	☐	☐	☐
Set57	☐	☐	☐	☐	☐	☐	☐
Set58	☐	☐	☐	☐	☐	☐	☐
Set59	☐	☐	☐	☐	☐	☐	☐
Set60	☐	☐	☐	☐	☐	☐	☐

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Set61	☐	☐	☐	☐	☐	☐	☐
Set62	☐	☐	☐	☐	☐	☐	☐
Set63	☐	☐	☐	☐	☐	☐	☐
Set64	☐	☐	☐	☐	☐	☐	☐
Set65	☐	☐	☐	☐	☐	☐	☐
Set66	☐	☐	☐	☐	☐	☐	☐
Set67	☐	☐	☐	☐	☐	☐	☐
Set68	☐	☐	☐	☐	☐	☐	☐
Set69	☐	☐	☐	☐	☐	☐	☐
Set70	☐	☐	☐	☐	☐	☐	☐
Set71	☐	☐	☐	☐	☐	☐	☐
Set72	☐	☐	☐	☐	☐	☐	☐
Set73	☐	☐	☐	☐	☐	☐	☐
Set74	☐	☐	☐	☐	☐	☐	☐
Set75	☐	☐	☐	☐	☐	☐	☐

	Worst regress			No change			Best improve
	-3	-2	-1	0	+1	+2	+3
Set76	☐	☐	☐	☐	☐	☐	☐
Set77	☐	☐	☐	☐	☐	☐	☐
Set78	☐	☐	☐	☐	☐	☐	☐
Set79	☐	☐	☐	☐	☐	☐	☐
Set80	☐	☐	☐	☐	☐	☐	☐
Set81	☐	☐	☐	☐	☐	☐	☐
Set82	☐	☐	☐	☐	☐	☐	☐
Set83	☐	☐	☐	☐	☐	☐	☐
Set84	☐	☐	☐	☐	☐	☐	☐
Set85	☐	☐	☐	☐	☐	☐	☐
Set86	☐	☐	☐	☐	☐	☐	☐
Set87	☐	☐	☐	☐	☐	☐	☐
Set88	☐	☐	☐	☐	☐	☐	☐
Set89	☐	☐	☐	☐	☐	☐	☐
Set90	☐	☐	☐	☐	☐	☐	☐

YOU NOW HAVE 1-MINUTE BREAK

Worst	No change	Best

	regress -3	-2	-1	0	+1	+2	improve +3
Set91	☐	☐	☐	☐	☐	☐	☐
Set92	☐	☐	☐	☐	☐	☐	☐
Set93	☐	☐	☐	☐	☐	☐	☐
Set94	☐	☐	☐	☐	☐	☐	☐
Set95	☐	☐	☐	☐	☐	☐	☐
Set96	☐	☐	☐	☐	☐	☐	☐
Set97	☐	☐	☐	☐	☐	☐	☐
Set98	☐	☐	☐	☐	☐	☐	☐
Set99	☐	☐	☐	☐	☐	☐	☐
Set100	☐	☐	☐	☐	☐	☐	☐
Set101	☐	☐	☐	☐	☐	☐	☐
Set102	☐	☐	☐	☐	☐	☐	☐
Set103	☐	☐	☐	☐	☐	☐	☐
Set104	☐	☐	☐	☐	☐	☐	☐
Set105	☐	☐	☐	☐	☐	☐	☐
	Worst			No change			Best

	regress -3	-2	-1	0	+1	+2	improve +3
Set106	☐	☐	☐	☐	☐	☐	☐
Set107	☐	☐	☐	☐	☐	☐	☐
Set108	☐	☐	☐	☐	☐	☐	☐
Set109	☐	☐	☐	☐	☐	☐	☐
Set110	☐	☐	☐	☐	☐	☐	☐
Set111	☐	☐	☐	☐	☐	☐	☐
Set112	☐	☐	☐	☐	☐	☐	☐
Set113	☐	☐	☐	☐	☐	☐	☐
Set114	☐	☐	☐	☐	☐	☐	☐
Set115	☐	☐	☐	☐	☐	☐	☐
Set116	☐	☐	☐	☐	☐	☐	☐
Set117	☐	☐	☐	☐	☐	☐	☐
Set118	☐	☐	☐	☐	☐	☐	☐
Set119	☐	☐	☐	☐	☐	☐	☐
Set120	☐	☐	☐	☐	☐	☐	☐
	Worst			No change			Best

	regress						improve
	-3	-2	-1	0	+1	+2	+3
Set121	☐	☐	☐	☐	☐	☐	☐
Set122	☐	☐	☐	☐	☐	☐	☐
Set123	☐	☐	☐	☐	☐	☐	☐
Set124	☐	☐	☐	☐	☐	☐	☐
Set125	☐	☐	☐	☐	☐	☐	☐
Set126	☐	☐	☐	☐	☐	☐	☐
Set127	☐	☐	☐	☐	☐	☐	☐
Set128	☐	☐	☐	☐	☐	☐	☐
Set129	☐	☐	☐	☐	☐	☐	☐
Set130	☐	☐	☐	☐	☐	☐	☐
Set131	☐	☐	☐	☐	☐	☐	☐
Set132	☐	☐	☐	☐	☐	☐	☐
Set133	☐	☐	☐	☐	☐	☐	☐
Set134	☐	☐	☐	☐	☐	☐	☐
Set135	☐	☐	☐	☐	☐	☐	☐

THANK YOU FOR YOUR PARTICIPATION